Asymmetric Invertible Threat: Learning Reversible Privacy Defense for Face Recognition

Jiabei Zhang[1,*] **Ziyuan Yang**[1,*] **Andrew Beng Jin Teoh**[2] **Yi Zhang**[1,†]

[1]School of Cyber Science and Engineering, Sichuan University, China
[2]School of Electrical and Electronic Engineering, Yonsei University, Republic of Korea

2023141530097@stu.scu.edu.cn cziyuanyang@gmail.com
bjteoh@yonsei.ac.kr yzhang@scu.edu.cn

[*] These authors contributed equally to this work. [†] Corresponding author.

Abstract

Face recognition systems are widely deployed in real-world applications, but they also raise privacy concerns due to unauthorized collection and misuse of facial data. Existing adversarial privacy protection methods rely on input-space perturbations to obfuscate identity information, yet their protection can degrade when adversaries learn restoration or purification mappings that partially invert the transformation. We study this setting as an asymmetric adversarial attack, in which reverse manipulation becomes feasible because existing defense paradigms do not control reversibility. To address this problem, we propose Asymmetric Reversible Face Protection (ARFP), a restoration-aware extension of personalized face cloaking that integrates privacy protection, keyed recovery, and tamper indication in a single framework. ARFP consists of three components: Key-Conditioned Manifold Binding, which ties the protection transformation to a user-provided key; Adversarial Restoration-Aware Training, which introduces a surrogate restoration adversary during training to improve robustness against evaluated inverse purification attacks; and Authorized Reversible Restoration, which supports recovery with the correct key while providing nonce-based tamper indication. Extensive experiments under the threat models considered in this work show that ARFP improves resistance to the evaluated restoration attacks while preserving authorized recovery utility. These results provide empirical evidence of key-sensitive recovery behavior and tamper awareness in the tested settings.[1]

1 Introduction

The widespread deployment of Face Recognition (FR) systems [19, 4, 22, 15] has transformed digital identity verification, supporting applications ranging from secure authentication to smart surveillance. However, it also raises substantial privacy concerns. Facial images shared online can be scraped to construct large-scale databases without user consent [20, 2], enabling misuse such as covert tracking, identity theft, and behavioral profiling [21, 12].

To mitigate such risks, adversarial privacy protection has emerged as a practical defense direction [20, 2]. Methods such asFawkes [20], LowKey [2], and Chameleon [3] perturb the input image to suppress unauthorized FR while preserving visual utility. Chameleon [3], in particular, introduces personalized privacy masks to improve efficiency and transferability. Despite these advances, existing

[1]The code has been released at https://github.com/Jacob-Zhang-cpu/ARFP.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

methods primarily operate as input-space cloaks and do not explicitly control whether the applied transformation can later be approximated, neutralized, or partially reversed [7, 10, 24].

Despite their practical effectiveness, these approaches rely on a one-sided protection assumption: once a perturbation or privacy mask is applied, the resulting image is implicitly treated as protected. This overlooks a fundamental asymmetry between intended access control and the model's actual capabilities. The forward transformation is meant to restrict identity inference, but no corresponding mechanism constrains reverse manipulation. Consequently, once restoration becomes feasible, it is not reserved for authorized users and may also be exploited by an adversary.

We study this vulnerability as an ***asymmetric adversarial attack***, in which an adversary learns or approximates an inverse mapping to reconstruct or weaken a privacy-preserving transformation without privileged access. Figure 1 shows this setting. The vulnerability arises because existing cloaking methods remain in the same model space used for training and evaluation. If the protection transformation can be approximated there, reverse manipulation can also be learned. The practical question is not only whether a cloak disrupts recognition, but whether reversal is restricted to authorized parties.

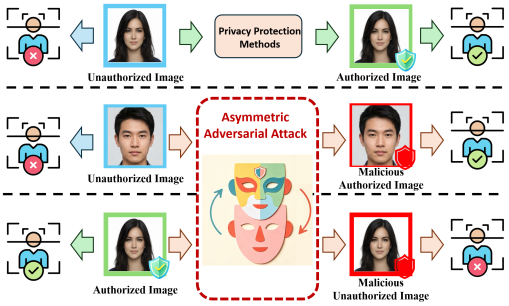

Figure 1: Illustration of the asymmetric adversarial attack.

Face privacy protection is therefore an *asymmetric, bidirectional problem*: a forward mapping generates a protected image for public release, whereas the reverse mapping should be available only under authorized conditions. This motivates key-conditioned protection and direct robustness training against restoration attacks.

To address this problem, we propose ***Asymmetric Reversible Face Protection (ARFP)***, which is best viewed as a restoration-aware extension of personalized face cloaking. ARFP integrates three modules: Key-Conditioned Manifold Binding (KMB), Adversarial Restoration-Aware Training (ARAT), and Authorized Reversible Restoration (ARR). KMB embeds a user-specific secret key into the mask generation process via affine modulation, producing a key-dependent transformation manifold. ARAT introduces a surrogate restoration network during training and optimizes the protector against this explicit inverse adversary. ARR supports key-guided recovery for authorized users while extracting a nonce as a fragile tamper-indication signal. Among these components, the central contribution is the restoration-aware training loop, which improves robustness to restoration attempts within the same protection pipeline. ARFP is model-agnostic and can be integrated with existing FR systems without retraining the recognizer.

The main contributions of this work are summarized as follows:

- We study restoration-based attacks against adversarial face cloaking in the asymmetric adversarial attack setting, showing that exposure of reversibility can weaken existing input-space protections.
- We present ARFP as a restoration-aware extension of personalized face cloaking. Its central technical contribution is an adversarial restoration-aware training loop, which improves robustness to restoration attempts within a key-conditioned protection pipeline.
- We integrate key-conditioned protection, authorized recovery, and nonce-based fragile-tamper indication into a single framework, and evaluate these components under the evaluated threat settings.

2 Related Work

Deep learning has substantially advanced FR performance [19, 4]. However, FR systems remain vulnerable to adversarial manipulation [21, 12]. Early digital cloaking methods, such as Fawkes [20], generate identity-shifting perturbations but require costly optimization and exhibit limited robustness to common image transformations. LowKey [2] improves black-box transferability through ensemble-based attacks, but still incurs substantial computation and may introduce visible artifacts. Physical-world methods, including Adv-Makeup [24] and AdvHat [13], target patch-based attacks and are less

suited to protecting digital images. Chameleon [3] further improves efficiency and transferability by generating personalized privacy masks with a conditional generative model.

Within this line of work, ARFP is more appropriately viewed as a restoration-aware extension of personalized face cloaking in the digital setting, rather than as a general solution to recoverable or verifiable privacy protection. The central distinction is the incorporation of a surrogate restoration adversary during training so that the learned protection remains more robust to the evaluated inverse or purification attacks, while still supporting keyed recovery and tamper indication.

Closely related to ARFP are key-conditioned or reversible face de-identification methods. Riddle [14] learns a latent encryptor that makes de-identification reversible given the correct key. Personalized invertible de-identification [1] disentangles identity and attribute representations to enable recovery. Password-conditioned anonymization [8] conditions a face identity transformer on a user-provided password. These methods provide credential-conditioned or reversible de-identification, but none trains against an explicit restoration adversary; their evaluations focus on protection against standard recognition models and recovery fidelity. In contrast, ARAT directly optimizes the protector against a surrogate restoration network, which is the central distinction of ARFP under the asymmetric adversarial setting.

3 Security Analysis and Empirical Insights

3.1 Threat Model

Adversary. In our threat model, the adversary can access protected facial images as well as publicly available face images from the web, but has no access to the user's secret key or the internal parameters of the protection mechanism. The adversary does not require knowledge of, or access to, the used FR model. Under such a setting, the adversary can readily collect large-scale facial data and perform identity inference without additional system-specific information. We additionally consider a stronger adversary that may obtain paired unprotected and protected face images and train a restoration network to invert the protection; this setting exceeds our primary threat model and is evaluated separately in Appendix B.

Defender. The defender operates without any special privileges over the FR system and does not modify or access its internal parameters. Under this setting, we identify the following defense goals:

1. Rejection of Unauthorized Inputs. Unauthorized face images should be rejected by the protected FR systems to reduce identity leakage from raw inputs.

2. Visual Fidelity Preservation. The protection process should preserve visual quality and usability for legitimate image sharing.

3. Identity Obfuscation. Protected images should suppress identity information cues against adversarial FR models while maintaining the intended asymmetric protection setting.

4. Authorized Recovery. The mechanism should support the recovery of identity-consistent images under the correct credential.

These four goals are defined with respect to the tested face-recognition models and the restoration attacks evaluated in this work, rather than as universal guarantees. Let $s_\phi(x_1, x_2) = \frac{\phi(x_1)^\top \phi(x_2)}{\|\phi(x_1)\| \, \|\phi(x_2)\|}$ denote the identity similarity under a fixed recognition model ϕ. Identity obfuscation requires that the protected image, as well as each protected image after the evaluated restoration attacks, yields low s_ϕ to the original subject. Authorized recovery requires that recovery with the correct key produces both low reconstruction distortion and high s_ϕ to the original image.

Existing methods mainly address the first two objectives. By contrast, our formulation is designed to address all four within a unified framework, with particular emphasis on the risk posed by inversion-oriented attacks under asymmetric access conditions.

3.2 Empirical Motivation

We conduct preliminary experiments to motivate our study by comparing representative prior work, Chameleon [3], with ARFP under inversion-oriented transformations. We use several domain-translation models to simulate restoration or inversion attempts. As shown in Figure 2(a), Chameleon

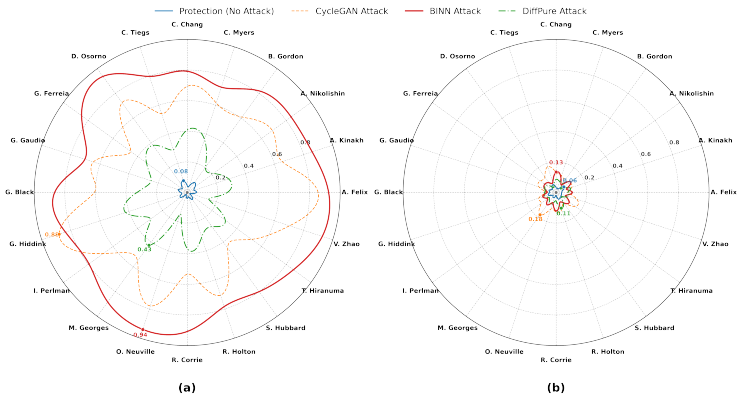

Figure 2: Feature Similarity under attacks (Lower is better). (a) Baseline fails against attacks. (b) ARFP maintains robust protection.

effectively rejects conventional recognition when such transformations are absent. However, under domain-transferred inputs, its protection degrades substantially for most evaluated inversion methods. In particular, unsupervised transformation models such as CycleGAN [28] and DiffPure [17] reduce the protection effect, while the supervised domain-transfer method Adaptive Reversal causes more pronounced degradation. These results suggest that, for the transformation families evaluated here, static obfuscation may not remain stable under distribution-altering but label-consistent mappings.

In contrast, Figure 2 (b) shows that ARFP maintains lower identity similarity under the same evaluated transformations. Its key-conditioned design appears more resistant to these inversion-oriented attacks in our empirical setting while preserving the asymmetric access structure intended by the framework. These preliminary results motivate the need for an asymmetric defense paradigm, though they should be interpreted within the scope of the evaluated attack models.

4 Method

4.1 Overview of the ARFP Framework

To preserve privacy and defend against inversion-oriented attacks, we propose the Asymmetric Reversible Face Protection (ARFP) framework, as shown in Figure 3. ARFP formulates face protection as a key-conditioned generative process with three coupled components. Key-Conditioned Manifold Binding (KMB) injects a user-specific key into the protection mapping so the generated perturbation depends on both image content and credential. Adversarial Restoration-Aware Training (ARAT) introduces a surrogate restoration adversary during optimization to improve resistance to inversion attempts. Authorized Reversible Restoration (ARR) enables credential-conditioned recovery and provides an auxiliary tamper signal based on a nonce. Together, these components define a protection pipeline with keyed conditioning, restoration-aware training, and recovery support under asymmetric access conditions.

4.2 Key-Conditioned Manifold Binding

To condition protection on user credentials, we introduce the KMB module. The conditional mask generator G is based on ResNet and receives an input image x, a secret key k, and a stochastic nonce m. The key k is a 256-bit binary vector with independent uniformly random bits, randomly generated and assigned to each user; the nonce m is a 64-bit binary vector generated identically. The two are concatenated and normalized to unit ℓ_2 norm to form a 320-dimensional conditioning input. Rather than using a fixed perturbation pattern, KMB modulates intermediate generator features with a condition embedding derived from (k, m), so that the resulting protection depends jointly on image content and the provided credential.

Let $x \in \mathbb{R}^{3\times H\times W}$ be the original image, and let $f_i \in \mathbb{R}^{C_i\times H_i\times W_i}$ denote the feature map at the i-th layer of the generator. We introduce a condition encoder E implemented as a multi-layer perceptron, which maps the key–nonce pair (k, m) to a latent embedding $e = E(k, m)$. KMB then recalibrates channel responses through learned affine modulation:

$$\text{KMB}(f_i, k, m) = \gamma_i(E(k, m)) \odot f_i + \beta_i(E(k, m)), \tag{1}$$

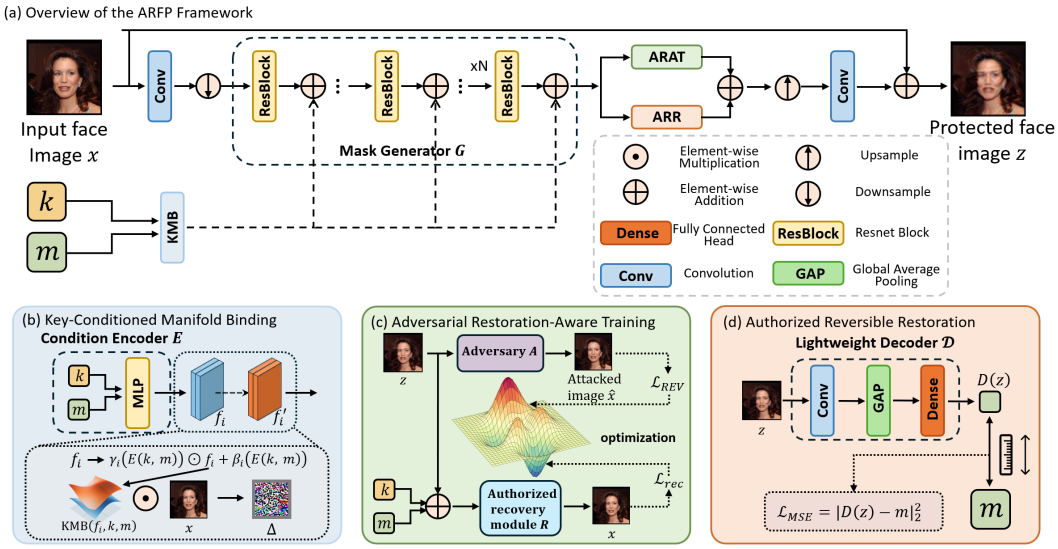

Figure 3: Overview of the proposed Asymmetric Reversible Face Protection (ARFP) framework. Our framework comprises three components: (b) KMB for key-conditioned feature modulation, (c) ARAT for robustness against restoration attacks, and (d) ARR for authorized reversible recovery.

where $\gamma_i(E(k,m)), \beta_i(E(k,m)) \in \mathbb{R}^{C_i}$ denote channel-wise scale and shift parameters that are broadcast over the spatial dimensions. The final protective perturbation $\Delta(x,k,m)$ is produced by aggregating these modulated features across the generator, and the protected image is defined as:

$$z = G(x,k,m) = x + \alpha \cdot \Delta(x,k,m), \tag{2}$$

where α controls the perturbation magnitude.

Under this construction, Δ is not a fixed additive mask but an input- and key-dependent transformation induced by the modulated generator features. In this sense, KMB makes protection mapping conditional on the supplied credential, consistent with the empirical key-sensitivity analysis presented later, rather than implying a general guarantee of irreversibility.

4.3 Adversarial Restoration-Aware Training

Standard adversarial training typically optimizes against a classifier. By contrast, ARAT formulates face privacy protection as a minimax optimization problem against an explicit inversion adversary. We write the objective as:

$$\min_{G,R} \max_{A} \mathbb{E}_{x,k,m} \Big[\|x - R(G(x,k,m),k,m)\|_1 - \lambda_{rev} \mathcal{L}_{REV}(x, A(G(x,k,m))) \Big], \tag{3}$$

where A denotes a surrogate restoration adversary network [5], and R is the authorized recovery module. For convenience, we denote the authorized reconstruction term by:

$$\mathcal{L}_{rec}(x,\hat{x}) = \|x - \hat{x}\|_1, \tag{4}$$

with $\hat{x} = R(G(x,k,m),k,m)$. The adversarial restoration loss is defined as:

$$\mathcal{L}_{REV}(x,\tilde{x}) = \|x - \tilde{x}\|_1 + \lambda_{l2} \|x - \tilde{x}\|_2^2, \tag{5}$$

where $\tilde{x} = A(G(x,k,m))$ and λ_{l2} controls the relative contribution of the ℓ_2 term. The coefficient λ_{rev} balances recovery fidelity and restoration resistance.

Under this formulation, A is optimized to minimize $\mathcal{L}_{REV}$ and thereby approximate an inverse mapping from protected images z to clean images x. In contrast, G and R are optimized so that authorized recovery remains feasible while the adversary's restoration becomes more difficult. In each training cycle, we alternate these updates: the surrogate adversary minimizes $\mathcal{L}_{REV}$ with the protector fixed, and the protector-side modules are then updated with the adversary fixed. This makes the optimization schedule explicit and removes ambiguity in the compact minimax notation.

The role of ARAT is therefore restoration-aware rather than purely classifier-aware. Its objective is to train the protection mapping against the specific restoration pressure induced by A, while preserving recoverability under the provided key.

4.4 Authorized Reversible Restoration

We next introduce the ARR module, which provides credential-conditioned recovery together with a lightweight tamper-indication signal. A nonce m is embedded during generation as an auxiliary conditioning variable. A lightweight decoder D is trained to recover this nonce from the protected image z:

$$\mathcal{L}_{MSE} = \|D(z) - m\|_2^2. \tag{6}$$

At inference time, the same decoder can be applied to z and, when needed, to the recovered image $\hat{x}$ to check whether the embedded signal remains consistent with the original nonce.

Consistency between the extracted nonce $\hat{m}$ and the original m serves as a practical indicator of significant modification. This branch is intended as fragile tamper indication or provenance awareness, not as cryptographic proof of ownership.

The total training objective is:

$$\mathcal{L}_{Total} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{id}\mathcal{L}_{id} + \lambda_{MSE}\mathcal{L}_{MSE} - \lambda_{rev}\mathcal{L}_{REV}, \tag{7}$$

where $\mathcal{L}_{id}$ is the identity loss, implemented by maximizing cosine distance in FR feature space to suppress identity matching. The hyperparameters λ_{rec}, λ_{id}, and λ_{MSE} control the relative weights of recovery fidelity, identity obfuscation, and nonce consistency, respectively. The term $-\lambda_{rev}\mathcal{L}_{REV}$ couples the protector to the surrogate adversary by penalizing solutions that are easy to restore.

In practice, once A is updated to reduce $\mathcal{L}_{REV}$, the protector-side modules are updated to minimize $\mathcal{L}_{Total}$ with A fixed. This alternating procedure clarifies how ARR interacts with ARAT during optimization.

4.5 Information-Theoretic Motivation

We provide a brief intuitive motivation for the key-conditioned design. Let $h = \phi(x)$ denote the identity embedding of a clean image x under a fixed face-recognition model ϕ, and let $z = G(x, k)$ be the protected output. An adversary that estimates h from z under squared error achieves a minimum error governed by the conditional uncertainty of h given z. Conditioning the protection on a high-entropy key is intended to reduce the identity information observable through z, making feature estimation harder on average for an adversary without the key. This is a qualitative design motivation rather than a formal security guarantee; the empirical resistance of ARFP to inversion attacks is established experimentally in Section 5 and Appendix B.

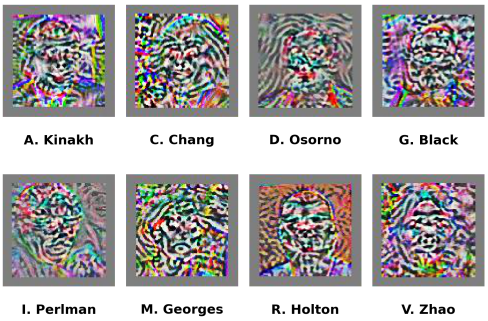

Figure 4: Visualization of rescaled protection masks. ARFP concentrates identity-specific perturbations on critical semantic regions.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate our framework on two benchmark datasets widely used in face privacy research: LFW [9] and FaceScrub [16]. Following prior work, we retain identities with at least 15 images and randomly split each identity's images into reference and probe sets. Detailed preprocessing operations are provided in Appendix A.

Face Recognition Models. To evaluate robustness and transferability, we use a diverse collection of publicly available pretrained FR models that span five architectures and four training datasets. These include EfficientNet (a non-ResNet architecture) as well as ResNet18/34/50/100 variants, trained on MS-Celeb-1M, Glint360K, VGG-Face2, and WebFace600K. Details are provided in Appendix A.

Table 1: Protection Success Rate (PSR %) of ARFP across eight black-box models.

Identity	PSR (%)								Statistics	
	EN-MC	RN50-GL	RN50-MC	RN50-VF	RN50-WF	RN18-MC	RN34-MC	RN100-MC	Mean	Std
A. Felix	100.00	100.00	93.33	86.67	100.00	93.33	86.67	100.00	95.00	5.27
B. Gordon	100.00	100.00	91.67	100.00	100.00	91.67	100.00	91.67	96.88	4.14
C. Myers	100.00	100.00	81.82	100.00	100.00	90.91	81.82	100.00	94.32	7.91
D. Osorno	100.00	92.31	100.00	92.31	100.00	92.31	84.62	100.00	95.19	5.56
G. Ferreia	100.00	100.00	100.00	92.86	100.00	92.86	92.86	100.00	97.32	3.57
I. Perlman	100.00	100.00	80.00	100.00	100.00	90.00	80.00	100.00	93.75	8.84
M. Georges	100.00	100.00	93.33	100.00	100.00	93.33	86.67	100.00	96.67	4.71
O. Neuville	91.67	100.00	100.00	100.00	91.67	100.00	91.67	100.00	96.88	4.14
R. Corrie	100.00	100.00	93.75	100.00	100.00	93.75	93.75	100.00	97.66	3.12
V. Zhao	100.00	100.00	90.91	100.00	100.00	90.91	81.82	100.00	95.45	6.57
Mean	**99.17**	**99.23**	**92.48**	**97.18**	**99.17**	**92.91**	**87.99**	**99.17**	**95.91**	**4.22**

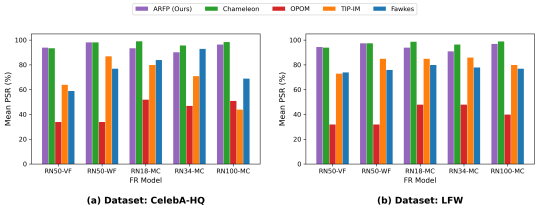

Figure 5: Protection effectiveness on LFW against unseen FR models.

Baselines. We compare ARFP against representative face-cloaking baselines under the relevant evaluation settings. For the unseen-model LFW comparison in Fig. 5, we report methods with directly comparable results under that protocol: Chameleon [3], OPOM [27], and TIP-IM [23]. For restoration-attack and image-quality evaluations, we additionally compare with Fawkes (High mode) [20] and Chameleon [3]. All baselines are run using their official implementations and recommended configurations.

Implementation Details. Comprehensive network and training hyperparameters are provided in Appendix A.

Evaluation Metric. Following prior work [3], we adopt the Protection Success Rate (PSR) to evaluate performance against unauthorized FR. PSR is defined as the complement of FR accuracy on the protected dataset:

$$\text{PSR} = 100\% - \text{FR Accuracy}. \tag{8}$$

A higher PSR indicates more effective suppression of matching to the ground-truth identity.

5.2 Protection Performance

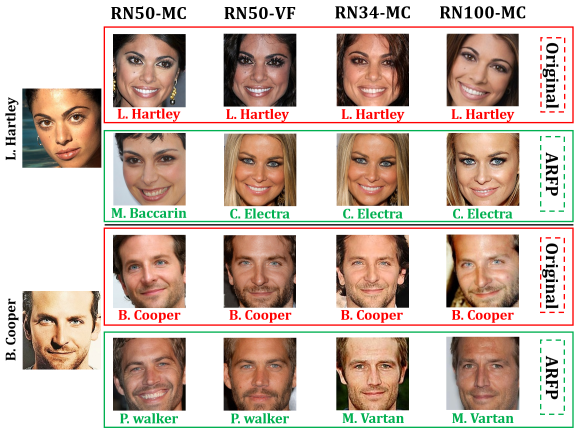

Figure 6: Cross-model retrieval on FaceScrub after protection.

To qualitatively assess the generated protection mechanisms, we visualize the learned masks and the protected images they produce. Unlike fixed or identity-agnostic masks, our approach produces adaptive perturbations for different faces, enabling more personalized protection.

Figure 4 illustrates the learned masks for selected identities. The masks exhibit distinct identity-dependent structures driven by the KMB module, suggesting that the protection adapts to subject-specific facial content.

To evaluate protection effectiveness and generalization across unseen evaluators, we compare ARFP with Chameleon [3], OPOM [27], and TIP-IM [23]. As shown in Fig. 5, ARFP achieves high unseen-model protection, with average PSR above 90%, and remains competitive with or better than the reported baselines across standard architectures.

Table 1 provides a finer-grained evaluation by reporting PSR for ten randomly selected identities against eight black-box FR models. These evaluators vary in architecture and pretraining source.

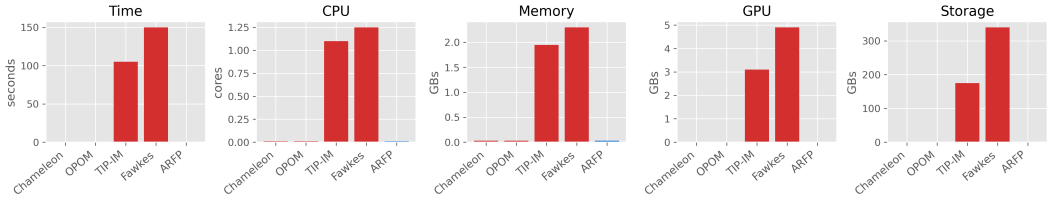

Figure 7: Protection efficiency analysis. ARFP uses a single forward pass, whereas optimization-based baselines incur a higher generation cost.

Table 2: Quantitative image quality comparison.

Method	PSNR (dB) ↑	SSIM ↑
Fawkes [20]	22.4	0.81
Chameleon [3]	28.5	0.89
ARFP (Ours)	**31.2**	**0.94**

Table 3: Full-scale restoration attack evaluation on the LFW test set. Mean PSR (%) $\pm$ variance.

Method	DiffPure	CycleGAN	BINN
Fawkes	23.1 ± 5.1	18.4 ± 4.2	10.1 ± 3.2
Chameleon	70.0 ± 8.0	32.0 ± 12.0	22.0 ± 8.0
ARFP (Ours)	$\mathbf{93.4} \pm 1.8$	$\mathbf{86.5} \pm 2.5$	$\mathbf{91.2} \pm 2.1$

ARFP achieves a mean PSR of 95.91%, with moderate variance across identities, indicating consistent protection performance across the tested models and subjects.

To qualitatively illustrate black-box transferability, we visualize retrieval results for unauthorized queries. As shown in Fig. 6, after applying ARFP, the retrieved identities are no longer the true matches.

We also evaluate computational cost, as shown in Fig. 7. Optimization-based approaches incur substantial overhead during protection generation. In contrast, ARFP requires only a single forward pass, achieving near real-time performance (≈ 0.05s) with low CPU and memory overhead, which is favorable for practical deployment.

5.3 Image Quality and Consistency

Preserving visual quality is an important objective for practical deployment. ARFP is designed to learn a transformation that maintains perceptual consistency while suppressing identity information. As illustrated in Fig. 8, the generated images retain natural appearance and exhibit stable visual characteristics across variations in pose and lighting.

Table **??** shows that ARFP also preserves image fidelity well, with a PSNR of 31.2 dB and an SSIM of 0.94. We next evaluate its robustness against restoration-oriented inversion attacks.



Figure 8: Visual consistency of ARFP across pose and lighting variations.

5.3.1 Robustness Experiment

We consider three restoration baselines: DiffPure [17], CycleGAN [28], and BINN [11]. Table **??** summarizes the performance under these attacks. ARFP maintains a mean PSR of 91.2% under BINN, the strongest paired attack in this evaluation. For the evaluated restoration families, these results indicate that the key-conditioned transformation makes recovery of identity-consistent clean images more difficult than for the compared baselines. We additionally evaluate a stronger white-box key-recovery attack, in which the adversary has full access to the generator and recoverer; even under this setting, gradient-based key optimization remains near the random baseline (Appendix B). Visualizations of adversarially attacked images are provided in the Appendix C.

Table 4: Access control and key sensitivity analysis, assessing the separation between correct-key and wrong-key recovery.

Access Condition	Key Error (HD)	PSNR ↑	SSIM ↑	Nonce BER ↓
Authorized (k_{true})	0 bits	**31.2**	**0.94**	**1.2%**
Minor Mistake	1 bit	14.3	0.38	48.9%
Moderate Mistake	16 bits	12.8	0.33	50.1%
Random Guess	$\approx$ 128 bits	12.1	0.31	50.4%

Table 5: Ablation of ARFP components. We report PSR before and after the Adaptive Reversal attack, together with SSIM.

ARAT	KMB	ARR	Clean PSR (%)	Attacked PSR (%)	SSIM
×	✓	✓	93.8	**24.5**	**0.98**
✓	×	✓	94.1	41.2	0.85
✓	✓	×	**96.5**	88.4	0.94
✓	✓	✓	95.9	**91.2**	0.94

5.4 Ablation Study

To evaluate the contribution of each module, we conduct the ablation study reported in Table 5. Here, Clean PSR denotes the defense success rate measured directly on the protected images. Attacked PSR denotes the protection success rate after the protected images are subjected to a BINN attack designed to remove the privacy perturbations. These two metrics quantify robustness under standard and adaptive evaluation settings.

The Effect of ARAT. Removing the surrogate restoration adversary leads to a large drop under adaptive reversal. While the model still achieves high Clean PSR (93.8%), Attacked PSR drops to 24.5%. This shows that restoration-aware training is important for robustness to the evaluated inversion attack.

The Effect of KMB. Without key conditioning, the model becomes a deterministic transformation that is easier to approximate from paired data. As shown in Table 5, removing the key reduces Attacked PSR to 41.2% and lowers reconstruction quality to SSIM 0.85, indicating that key-dependent modulation contributes substantially to restoration resistance.

The Effect of ARR. Removing the nonce branch slightly improves Clean PSR (96.5%) but reduces Attacked PSR under purification (88.4% versus 91.2%). This suggests that the auxiliary nonce regularization stabilizes the learned transformation and modestly improves robustness in the evaluated setting.

5.5 Access Control

As shown in Table 4, correct-key access enables faithful recovery with high structural consistency (SSIM 0.94), whereas incorrect-key access leads to severe reconstruction failure (SSIM < 0.4). Meanwhile, the nonce branch rejects wrong-key outputs with a Bit Error Rate (BER) close to that of random guessing. Table 4 also shows that the recovery degradation saturates after a moderate number of incorrect bits (16-bit and 128-bit errors yield similar SSIM), whereas the present experiments only vary the Hamming error within a fixed 256-bit key and do not establish an effective security level associated with the nominal key length. These results indicate strong empirical key sensitivity for authorized recovery under the evaluated conditions. In addition, ARFP supports nonce-based integrity-aware recovery as a lightweight tamper-indication mechanism during restoration.

Table 6: Nonce BER under benign and malicious transformations.

Condition	BER (%)
Raw	1.2
JPEG Q=95	3.4
JPEG Q=75	4.5
JPEG Q=50	7.1
JPEG Q=30	9.1
Resize 0.50×	5.4
JPEG Q=75 ×2	6.0
Splicing [6]	48.3
BINN [11]	49.1
DiffPure [17]	50.2

We further evaluate the robustness of integrity-aware recovery under benign and malicious conditions. Benign transformations include raw images, multiple JPEG qualities, resizing, and repeated compression. Malicious scenarios include local tampering through splicing [6], as well as two restoration-oriented attacks: BINN [11] and DiffPure [17]. As shown in Table 6, the evaluated benign transformations produce BER values well within the low range, while the malicious operations produce BER near the random-guessing level. This separation supports the nonce branch as a practical indicator of tampering or restoration failure, rather than as a general verification mechanism. The full set of transformations, together with a discussion of the operational detection threshold, is provided in Appendix B.

5.6 Hyperparameter Selection

We further analyze the sensitivity of ARFP to the perturbation magnitude factor α, which controls the trade-off between protection strength and visual quality. As illustrated in Figure 9, varying α from 0.01 to 0.15 reveals three regimes. For small perturbations ($\alpha < 0.03$), visual quality remains high (PSNR > 35 dB), but protection against strong FR models is insufficient. Around $\alpha = 0.05$, the curve exhibits a knee point: further increases produce smaller gains in protection while degrading perceptual quality more noticeably. We therefore select $\alpha = 0.05$ as the operating point used in the main experiments. For larger perturbations ($\alpha > 0.08$), protection continues to increase, but visible artifacts become more pronounced and PSNR declines. A finer-grained sweep that additionally reports SSIM, LPIPS, and the protected-image cosine similarity at each α is provided in Appendix B.

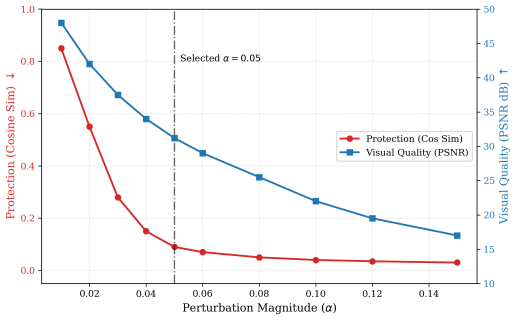

Figure 9: The impact of α.

6 Conclusion

In this paper, we identify the lack of controllable reversibility in existing face privacy protection methods, which exposes them to asymmetric adversarial attacks. To address this, we introduce an attack paradigm and propose the ARFP framework, formulating face privacy as an asymmetric bidirectional problem. ARFP integrates key-conditioned generation, restoration-aware adversarial training, and authorized recovery, improving empirical robustness against the evaluated restoration attacks while preserving reliable recovery. In addition, a nonce-based mechanism provides a lightweight signal for tamper awareness. Extensive experiments demonstrate that ARFP achieves strong empirical robustness, high visual fidelity, and efficient deployment. Future work includes extending the framework to broader modalities and exploring stronger theoretical guarantees for controllable reversibility.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under 62271335, in part by the Sichuan Science and Technology Program under Grant 2025ZNSFSC0470, and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NO. RS-2026-25587481).

References

[1] Jingyi Cao, Bo Liu, Yunqian Wen, Rui Xie, and Ling Song. Personalized and invertible face de-identification by disentangled identity information manipul ation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3334–3342, 2021.

[2] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P. Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social

media users from facial recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=hJmtwocEqzc.

[3] Ka-Ho Chow, Sihao Hu, Tiansheng Huang, and Ling Liu. Personalized privacy protection mask against unauthorized facial recognition. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXII*, volume 15140 of *Lecture Notes in Computer Science*, pages 434–450. Springer, 2025. ISBN 978-3-031-73006-1. URL https://doi.org/10.1007/978-3-031-73007-8_24.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00482. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

[5] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. 2015. URL http://arxiv.org/abs/1410.8516.

[6] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 422–426, 2013. doi: 10.1109/ChinaSIP.2013.6625374.

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

[8] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Alexander G Schwing. Password-conditioned anonymization and deanonymization with face identity transformers. In *European Conference on Computer Vision*, pages 727–743. Springer, 2020.

[9] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. Technical report, Marseille, France, October 2008. URL https://inria.hal.science/inria-00321923.

[10] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=iAmZUo0DxC0.

[11] Jörn-Henrik Jacobsen, Arnold W. M. Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HJsjkMb0Z.

[12] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection – a game theory perspective. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face ID system. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 819–826. IEEE, 2020. doi: 10.1109/ICPR48806.2021.9412236. URL https://doi.org/10.1109/ICPR48806.2021.9412236.

[14] Dongze Li, Wei Wang, Kang Zhao, Jing Dong, and Tieniu Tan. Riddle: Reversible and diversified de-identification with latent encryptor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8093–8102, 2023.

[15] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14225–14234.

Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01400. URL https://openaccess.thecvf.com/content/CVPR2021/html/Meng_MagFace_A_Universal_Representation_for_Face_Recognition_and_Quality_Assessment_CVPR_2021_paper.html.

[16] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 343–347, 2014. doi: 10.1109/ICIP.2014.7025068.

[17] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:248811081.

[18] Olivier Rioul. A simple proof of the entropy-power inequality via properties of mutual information. pages 46–50, 2007. doi: 10.1109/ISIT.2007.4557202. URL https://doi.org/10.1109/ISIT.2007.4557202.

[19] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[20] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In Srdjan Capkun and Franziska Roesner, editors, *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, pages 1589–1604. USENIX Association, 2020. URL https://www.usenix.org/conference/usenixsecurity20/presentation/shan.

[21] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi, editors, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 1528–1540. ACM, 2016.

[22] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5265–5274. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00552. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_CosFace_Large_Margin_CVPR_2018_paper.html.

[23] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3897–3907, October 2021.

[24] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1252–1258. ijcai.org, 2021. doi: 10.24963/IJCAI.2021/173. URL https://doi.org/10.24963/ijcai.2021/173.

[25] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503, 2016.

[26] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] Yaoyao Zhong and Weihong Deng. OPOM: customized invisible cloak towards face privacy protection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3590–3603, 2023. doi: 10.1109/TPAMI.2022.3175602. URL https://doi.org/10.1109/TPAMI.2022.3175602.

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

A Experimental Setup Details

A.1 Datasets and Protocols

(1) LFW: Represents an unconstrained recognition scenario with variations in pose, lighting, and expression. Following the protocol in [3, 2], we filter identities with at least 15 images to simulate users who actively post photos online. For each identity, we randomly split images into a reference set (for protection generation) and a probe set (for evaluation), ensuring no data leakage.

(2) FaceScrub: Used for identity-level robustness and generalization analysis. Its broad identity coverage supports rigorous evaluation of protection transferability across subjects.

All facial images in both datasets are pre-processed using MTCNN [25] for detection and alignment, then resized to 112×112 resolution, and finally normalized to the $[-1, 1]$ standard range before being fed into the protection network.

A.2 Face Recognition Models

To comprehensively evaluate the robustness and transferability of our protection, we employ the same diverse collection of publicly available pretrained FR models as in Chameleon [3], detailed in Table 7. These models cover a wide range of architectures and training subsets.

Table 7: The collection of publicly available pretrained FR models.

Model ID	Neural Arch.	Training Dataset	FR Acc.
EN-MC	EfficientNet	MS-Celeb-1M	94.94%
RN50-GL	ResNet50	Glint360K	96.68%
RN50-MC	ResNet50	MS-Celeb-1M	89.25%
RN50-VF	ResNet50	VGG-Face2	94.30%
RN50-WF	ResNet50	WebFace600K	96.72%
RN18-MC	ResNet18	MS-Celeb-1M	82.64%
RN34-MC	ResNet34	MS-Celeb-1M	84.49%
RN100-MC	ResNet100	MS-Celeb-1M	91.85%

A.3 Implementation Details

All experiments are implemented in TensorFlow and executed on a single NVIDIA RTX 4060 GPU. The protection generator follows the ResNet-based key-conditioned architecture described in the main method section, and the condition encoder is implemented as a lightweight MLP. Following the optimization protocol in Chameleon [3], we use the Adam optimizer with a learning rate of 2×10^{-4} and a batch size of 16. The balancing hyperparameters are set to $\lambda_{rec} = 10$, $\lambda_{id} = 5$, and $\lambda_{rev} = 1.0$, which were found empirically to provide stable convergence.

To model adaptive restoration attacks, we train a separate reversal network on the clean LFW training split until convergence and use it as the surrogate adversary in evaluation. The secret key k and nonce m are encoded as 256-bit and 64-bit binary vectors, respectively. The identity loss $\mathcal{L}_{id}$ is implemented as a cosine embedding loss over intermediate features extracted by a pretrained FR network. For paired inverse-attacker training, the surrogate adversary uses a batch size of 32, an initial learning rate of 10^{-4} decayed by 0.95 every 10 epochs, and paired samples generated from the underlying dataset.

For the white-box key-recovery attack, the key is optimized as a continuous relaxation with sigmoid binarization using the Adam optimizer. We test random initialization, near-key initialization (a single flipped bit), and partial key leakage (50% of the bits fixed to ground truth).

B Security Analysis under Stronger Adversaries

White-box key recovery attack. We evaluate a stronger adaptive adversary than the one assumed in our primary threat model. The adversary is given paired clean/protected samples (x, z), the

nonce m, and full access to the generator G and recoverer R, and attempts to recover the 256-bit key via gradient-based optimization over up to $10{,}000$ iterations, testing both reconstruction and identity-matching objectives. Table 8 reports representative results.

Table 8: White-box key recovery attack. Key accuracy is the fraction of correctly recovered bits over the full 256-bit key.

Attack Setting	**Key Accuracy**	**Recovery SSIM**	**PSR**
Random guessing (lower bound)	0.500	0.307	95.9%
Gradient attack, 10,000 iters	0.516	0.317	93.3%
Partial key leak, 50% bits	0.766	0.327	91.4%
True key (authorized upper bound)	1.000	0.941	5.1%

Even with full knowledge of the model, gradient optimization recovers the key only marginally above the random baseline, and recovery remains in the wrong-key regime. When 50% of the key bits are known, the remaining unknown bits are not recovered above the random baseline under the evaluated optimization protocol. We interpret this as empirical resistance under the evaluated white-box attack rather than as proof that key recovery is impossible.

Perceptual quality. Table **??** in the main text reports PSNR, SSIM, LPIPS, and FID measured with respect to the original unprotected image. LPIPS uses an AlexNet backbone; FID uses InceptionV3 pool3 features with standard preprocessing, computed over the evaluation images used in the main experiments.

Extended perturbation-magnitude sweep. Table 9 extends the analysis in Figure 9 with SSIM, LPIPS, and the protected-image cosine similarity at each perturbation magnitude.

Table 9: Extended perturbation-magnitude sweep. PSNR and CosSim are from the existing evaluation; SSIM and LPIPS are supplemented.

α	**PSNR (dB)**	**SSIM**	**LPIPS**	**CosSim**
0.01	48.0	0.992	0.039	0.85
0.03	37.5	0.968	0.061	0.28
0.05	31.2	0.936	0.074	0.09
0.08	25.5	0.896	0.098	0.05
0.10	22.0	0.873	0.118	0.04
0.15	17.0	0.827	0.148	0.03

Extended nonce robustness. Table 10 reports the nonce BER under a broader set of benign and malicious transformations. Benign operations produce BER at most 9.1%, whereas all malicious operations produce BER near 50%. With a detection threshold at a moderate BER, benign and malicious operations are well separated under the evaluated conditions.

Table 10: Nonce BER under extended benign and malicious transformations.

Transformation	**BER (%)**	**Transformation**	**BER (%)**
None (raw)	1.2	JPEG Q=50 + Resize 0.5×	8.8
JPEG Q=95	3.4	JPEG Q=75 ×2	6.0
JPEG Q=85	4.0	JPEG Q=75 ×3	7.6
JPEG Q=75	4.5	Splicing [6]	48.3
JPEG Q=60	5.9	BINN [11]	49.1
JPEG Q=50	7.1	DiffPure [17]	50.2
JPEG Q=30	9.1		
Resize 0.75×	3.7		
Resize 0.50×	5.4		
JPEG Q=75 + Resize 0.75×	6.5		

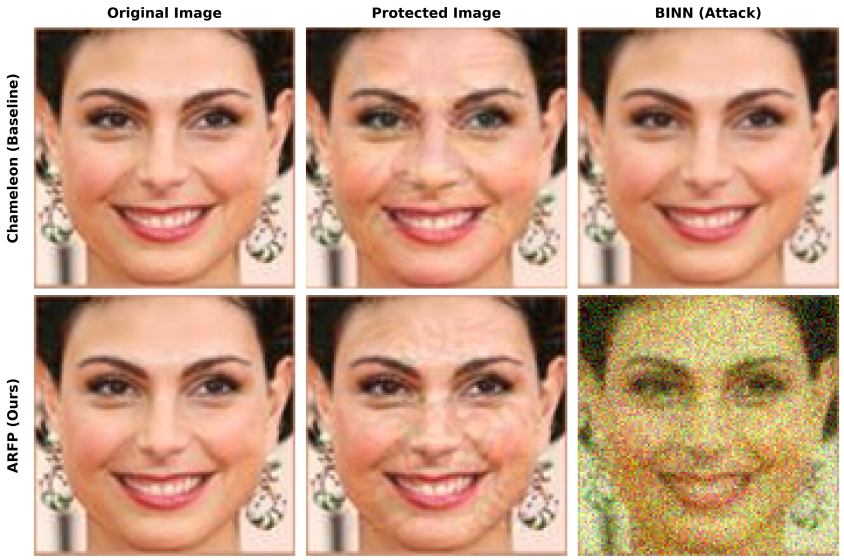

Figure 10: Qualitative purification examples. The evaluated attack procedure recovers more identity-consistent structure from Chameleon outputs (Top), whereas ARFP outputs (Bottom) remain substantially distorted under the same procedure.

C Qualitative purification examples

As shown in Fig. 10, the evaluated purification procedure produces visually more identity-consistent outputs for Chameleon than for ARFP. Under the same attack setting, the ARFP examples remain heavily distorted, indicating weaker recovery quality for the attacker in this qualitative comparison. These examples should be interpreted as illustrative evidence for the evaluated procedure rather than as a general statement about all possible restoration attacks.

D Discussion and Limitations

While ARFP demonstrates strong protection capabilities, there are several aspects worth discussing regarding its practical deployment.

Computational Overhead. Compared to traditional additive noise attacks like Fawkes, our method involves a forward pass through an adaptive reversal network, which is slightly more computationally intensive but still efficient. The primary cost lies in the training phase, where the adversarial loop requires oscillating updates. However, once trained, the generation of personalized protection is highly efficient due to its feed-forward design, making it feasible for real-time user-end applications.

Physical World Robustness. Our current evaluation focuses on the digital domain. In real-world scenarios where protected images might be printed and recaptured (print-scan attack), the high-frequency components of the adversarial perturbation and the fragile integrity indicator (Nonce) might be degraded. Enhancing the robustness of both the protection and the watermark against analog transformations remains a direction for future work.

Key Management. The security of our reversibility relies entirely on the secrecy of the user key k_u. This design choice, while structurally sound, introduces a key management burden. If a user loses their key, the original image cannot be mathematically recovered from the protected version. This intentional "failure" ensures that no backdoor exists for unauthorized parties, but it requires users to securely store their keys.